Name: _____

GCE Stats section 7

Correlation and linear regression

practice questions from GCE Maths part 1

Date:

Question parts in red cover content in the GCE Statistics specification but outside this section of the scheme of work

Time:

Total marks available:

Total marks achieved: _____

Questions

Q1.

A clothes shop manager records the weekly sales figures, £ s, and the average weekly temperature, t °C, for 6 weeks during the summer. The sales figures were coded so that $W=\frac{S}{1000}$

The data are summarised as follows

$$S_{ww} = 50 \quad \sum wt = 784 \quad \sum t^2 = 2435 \quad \sum t = 119 \quad \sum w = 42$$

(a) Find S_{wt} and S_{tt}

(3)

(b) Write down the value of S_{ss} and the value of S_{st}

(2)

(c) Find the product moment correlation coefficient between s and t.

(2)

The manager of the clothes shop believes that a linear regression model may be appropriate to describe these data.

(d) State, giving a reason, whether or not your value of the correlation coefficient supports the manager's belief.

(1)

(e) Find the equation of the regression line of w on t, giving your answer in the form $w = a + bt$

(3)

(f) Hence find the equation of the regression line of s on t, giving your answer in the form $s = c + dt$, where c and d are correct to 3 significant figures.

(2)

(g) Using your equation in part (f), interpret the effect of a 1°C increase in average weekly temperature on weekly sales during the summer.

(1)

(Total for question = 14 marks)

Q2.

A biologist is studying the behaviour of bees in a hive. Once a bee has located a source of food, it returns to the hive and performs a dance to indicate to the other bees how far away the source of the food is. The dance consists of a series of wiggles. The biologist records the distance, d metres, of the food source from the hive and the average number of wiggles, w, in the dance.

Distance, d m	30	50	80	100	150	400	500	650
Average number of wiggles, w	0.725	1.210	1.775	2.250	3.518	6.382	8.185	9.555

[You may use $\sum w = 33.6$ $\sum dw$ 13 833 $S_{dd} = 394\,600$ $S_{ww} = 80.481$ (to 3 decimal places)]

(a) Show that $S_{dw} = 5601$

(2)

(b) State, giving a reason, which is the response variable.

(1)

(c) Calculate the product moment correlation coefficient for these data.

(2)

(d) Calculate the equation of the regression line of w on d, giving your answer in the form $w = a + bd$

(4)

A new source of food is located 350 m from the hive.

(e) (i) Use your regression equation to estimate the average number of wiggles in the corresponding dance.

(ii) Comment, giving a reason, on the reliability of your estimate.

(2)

(Total for question = 11 marks)

Q3.

A meteorologist believes that there is a relationship between the height above sea level, h m, and the air temperature, t °C. Data is collected at the same time from 9 different places on the same mountain. The data is summarised in the table below.

h	1400	1100	260	840	900	550	1230	100	770
t	3	10	20	9	10	13	5	24	16

[You may assume that $\sum h = 7150$, $\sum t = 110$, $\sum h^2 = 7171500$, $\sum t^2 = 1716$, $\sum th = 64\,980$ and $S_{tt} = 371.56$]

(a) Calculate S_{th} and S_{hh}. Give your answers to 3 significant figures.

(3)

(b) Calculate the product moment correlation coefficient for this data.

(2)

(c) State whether or not your value supports the use of a regression equation to predict the air temperature at different heights on this mountain. Give a reason for your answer.

(1)

(d) Find the equation of the regression line of t on h giving your answer in the form $t = a + bh$.

(4)

(e) Interpret the value of b.

(1)

(f) Estimate the difference in air temperature between a height of 500 m and a height of 1000 m.

(2)

(Total 13 marks)

Q4.

Sammy is studying the number of units of gas, g, and the number of units of electricity, e, used in her house each week. A random sample of 10 weeks use was recorded and the data for each week were coded so that $x = \frac{g - 60}{4}$ and $y = \frac{e}{10}$. The results for the coded data are summarised below

$$\sum x = 48.0 \quad \sum y = 58.0 \quad S_{xx} = 312.1 \quad S_{yy} = 2.10 \quad S_{xy} = 18.35$$

(a) Find the equation of the regression line of y on x in the form $y = a + bx$.

Give the values of a and b correct to 3 significant figures.

(4)

(b) Hence find the equation of the regression line of e on g in the form $e = c + dg$.

Give the values of c and d correct to 2 significant figures.

(4)

(c) Use your regression equation to estimate the number of units of electricity used in a week when
100 units of gas were used.

(2)

(Total 10 marks)

Q5.

A teacher asked a random sample of 10 students to record the number of hours of television, t, they watched in the week before their mock exam. She then calculated their grade, g, in their mock exam.
The results are summarised as follows.

$$\sum t = 258 \quad \sum t^2 = 8702 \quad \sum g = 63.6 \quad S_{gg} = 7.864 \quad \sum gt = 1550.2$$

(a) Find S_{tt} and S_{gt}

(3)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between t and g.

(2)

The teacher also recorded the number of hours of revision, v, these 10 students completed during the week before their mock exam. The correlation coefficient between t and v was -0.753

(c) Describe, giving a reason, the nature of the correlation you would expect to find between v and g.

(2)

(Total 7 marks)

Q6.

On a particular day the height above sea level, x metres, and the mid-day temperature, y ºC, were recorded in 8 north European towns. These data are summarised below

$$S_{xx} = 3\,535\,237.5 \quad \sum y = 181 \quad \sum y^2 = 4305 \quad S_{xy} = -23\,726.25$$

(a) Find S_{yy}

(2)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient for these data.

(2)

(c) Give an interpretation of your coefficient.

(1)

A student thought that the calculations would be simpler if the height above sea level, h, was measured in kilometres and used the variable $h = \frac{x}{1000}$ instead of x.

(d) Write down the value of S_{hh}

(1)

(e) Write down the value of the correlation coefficient between h and y.

(1)

(Total 7 marks)

Q7.

A random sample of 50 salmon was caught by a scientist. He recorded the length l cm and weight w kg of each salmon.

The following summary statistics were calculated from these data.

$$\sum l = 4027 \quad \sum l^2 = 327\,754.5 \quad \sum w = 357.1 \quad \sum lw = 29\,330.5 \quad S_{ww} = 289.6$$

(a) Find S_{ll} and S_{lw}

(3)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between l and w.

(2)

(c) Give an interpretation of your coefficient.

(1)

(Total 6 marks)

Q8.

Gary compared the total attendance, x, at home matches and the total number of goals, y, scored at home during a season for each of 12 football teams playing in a league. He correctly calculated:

$$S_{xx} = 1022500 \quad S_{yy} = 130.9 \quad S_{xy} = 8825$$

(a) Calculate the product moment correlation coefficient for these data.

(2)

(b) Interpret the value of the correlation coefficient.

(1)

Helen was given the same data to analyse. In view of the large numbers involved she decided to divide the attendance figures by 100. She then calculated the product moment correlation coefficient between $\frac{x}{100}$ and y.

(c) Write down the value Helen should have obtained.

(1)

(Total 4 marks)

Q9.

An estate agent recorded the price per square metre, p £/m^2, for 7 two–bedroom houses.

He then coded the data using the coding $q = \frac{p - a}{b}$, where a and b are positive constants.

His results are shown in the table below.

p	1840	1848	1830	1824	1819	1834	1850
q	4.0	4.8	3.0	2.4	1.9	3.4	5.0

(a) Find the value of a and the value of b

(2)

The estate agent also recorded the distance, d km, of each house from the nearest train station. The results are summarised below.

$$S_{dd} = 1.02 \qquad S_{qq} = 8.22 \qquad S_{dq} = -2.17$$

(b) Calculate the product moment correlation coefficient between d and q

(2)

(c) Write down the value of the product moment correlation coefficient between d and p

(1)

The estate agent records the price and size of 2 additional two–bedroom houses, H and J.

House	Price (£)	Size (m^2)
H	156 400	85
J	172 900	95

(d) Suggest which house is most likely to be closer to a train station. Justify your answer.

(3)

(Total for question = 8 marks)

Q10.

A bank reviews its customer records at the end of each month to find out how many customers have become unemployed, u, and how many have had their house repossessed, h, during that month. The bank codes the data using variables $x = \frac{u - 100}{3}$ and $y = \frac{h - 20}{7}$

The results for the 12 months of 2009 are summarised below.

$$\sum x = 477 \quad S_{xx} = 5606.25 \quad \sum y = 480 \quad S_{yy} = 4244 \quad \sum xy = 23\,070$$

(a) Calculate the value of the product moment correlation coefficient for x and y.

(3)

(b) Write down the product moment correlation coefficient for u and h.

(1)

The bank claims that an increase in unemployment among its customers is associated with an increase in house repossessions.

(c) State, with a reason, whether or not the bank's claim is supported by these data.

(2)

(Total 6 marks)

Q11.

Before going on holiday to *Seapron*, Tania records the weekly rainfall (x mm) at *Seapron* for 8 weeks during the summer. Her results are summarised as

$$\sum x = 86.8 \qquad \sum x^2 = 985.88$$

(a) Find the standard deviation, σ_x, for these data.

(3)

Tania also records the number of hours of sunshine (y hours) per week at *Seapron* for these 8 weeks and obtains the following

$$\bar{y} = 58 \qquad \sigma_y = 9.461 \text{ (correct to 4 significant figures)} \qquad \sum xy = 4900.5$$

(b) Show that $S_{yy} = 716$ (correct to 3 significant figures)

(1)

(c) Find S_{xy}

(2)

(d) Calculate the product moment correlation coefficient, r, for these data.

(2)

During Tania's week-long holiday at *Seapron* there are 14 mm of rain and 70 hours of sunshine.

(e) State, giving a reason, what the effect of adding this information to the above data would be on the value of the product moment correlation coefficient.

(2)

(Total for question = 10 marks)

Q12.

The table shows data on the number of visitors to the UK in a month, v (1000s), and the amount of money they spent, m (£ millions), for each of 8 months.

Number of visitors v (1000s)	2450	2480	2540	2420	2350	2290	2400	2460
Amount of money spent m (£ millions)	1370	1350	1400	1330	1270	1210	1330	1350

You may use

$S_{vv} = 42587.5$ $S_{vm} = 31512.5$ $S_{mm} = 25187.5$ $\sum v = 19390$ $\sum m = 10610$

(a) Find the product moment correlation coefficient between m and v.

(2)

(b) Give a reason to support fitting a regression model of the form $m = a + bv$ to these data.

(1)

(c) Find the value of b correct to 3 decimal places.

(2)

(d) Find the equation of the regression line of m on v.

(2)

(e) Interpret your value of b.

(2)

(f) Use your answer to part (d) to estimate the amount of money spent when the number of visitors to the UK in a month is 2 500 000

(2)

(g) Comment on the reliability of your estimate in part (f). Give a reason for your answer.

(2)

(Total 13 marks)

Q13.

A large company is analysing how much money it spends on paper in its offices every year. The number of employees, x, and the amount of money spent on paper, p (£ hundreds), in 8 randomly selected offices are given in the table below.

x	8	9	12	14	7	3	16	19
p (£ hundreds)	40.5	36.1	30.4	39.4	32.6	31.1	43.4	45.7

(You may use $\sum x^2 = 1160$ $\sum p = 299.2$ $\sum p^2 = 11\ 422$ $\sum xp = 3449.5$

(a) Show that $S_{pp} = 231.92$ and find the value of S_{xx} and the value of S_{xp}

(5)

(b) Calculate the product moment correlation coefficient between x and p.

(2)

The equation of the regression line of p on x is given in the form $p = a + bx$.

(c) Show that, to 3 significant figures, $b = 0.824$ and find the value of a.

(4)

(d) Estimate the amount of money spent on paper in an office with 10 employees.

(2)

(e) Explain the effect each additional employee has on the amount of money spent on paper.

(1)

Later the company realised it had made a mistake in adding up its costs, p. The true costs were actually half of the values recorded. The product moment correlation coefficient and the equation of the linear regression line are recalculated using this information.

(f) Write down the new value of

(i) the product moment correlation coefficient,
(ii) the gradient of the regression line.

(2)

(Total 16 marks)

Q14.

A biologist is comparing the intervals (m seconds) between the mating calls of a certain species of tree frog and the surrounding temperature (t ºC). The following results were obtained.

t °C	8	13	14	15	15	20	25	30
m secs	6.5	4.5	6	5	4	3	2	1

(You may use $\sum tm = 469.5$, $S_{tt} = 354$, $S_{mm} = 25.5$)

(a) Show that $S_{tm} = -90.5$

(4)

(b) Find the equation of the regression line of m on t giving your answer in the form $m = a + bt$.

(4)

(c) Use your regression line to estimate the time interval between mating calls when the surrounding temperature is 10 ºC.

(1)

(d) Comment on the reliability of this estimate, giving a reason for your answer.

(1)

(Total 10 marks)

Q15.

A scientist is researching whether or not birds of prey exposed to pollutants lay eggs with thinner shells. He collects a random sample of egg shells from each of 6 different nests and tests for pollutant level, p, and measures the thinning of the shell, t. The results are shown in the table below.

p	3	8	30	25	15	12
t	1	3	9	10	5	6

[You may use $\sum p^2 = 1967$ and $\sum pt = 694$]

(a) Draw a scatter diagram on the axes on page 7 to represent these data.

(2)

(b) Explain why a linear regression model may be appropriate to describe the relationship between p and t.

(1)

(c) Calculate the value of S_{pt} and the value of S_{pp}.

(4)

(d) Find the equation of the regression line of t on p, giving your answer in the form $t = a + bp$.

(4)

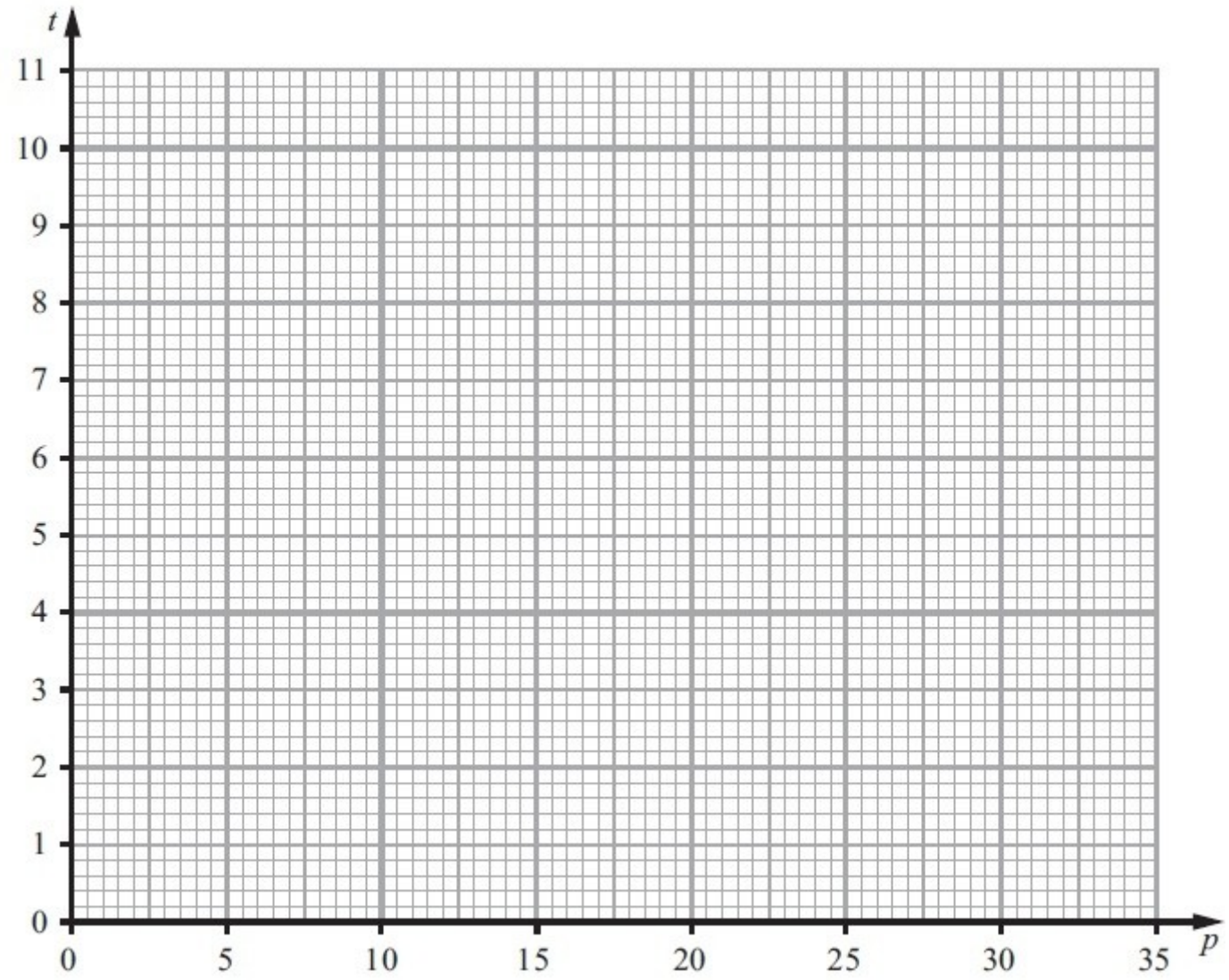
(e) Plot the point $(\overline{p}, \overline{t})$ and draw the regression line on your scatter diagram.

(2)

The scientist reviews similar studies and finds that pollutant levels above 16 are likely to result in the death of a chick soon after hatching.

(f) Estimate the minimum thinning of the shell that is likely to result in the death of a chick.

(2)



(Total 15 marks)

Q16.

Statistical models can provide a cheap and quick way to describe a real world situation.

(a) Give two other reasons why statistical models are used.

(2)

A scientist wants to develop a model to describe the relationship between the average daily
temperature, x °C, and her household's daily energy consumption, y kWh, in winter.

A random sample of the average daily temperature and her household's daily energy
consumption are taken from 10 winter days and shown in the table.

x	−0.4	−0.2	0.3	0.8	1.1	1.4	1.8	2.1	2.5	2.6
y	28	30	26	25	26	27	26	24	22	21

[You may use $\sum x^2 = 24.76$ $\sum y = 255$ $\sum xy = 283.8$ $S_{xx} = 10.36$]

(b) Find S_{xy} for these data.

(3)

(c) Find the equation of the regression line of y on x in the form $y = a + bx$

Give the value of a and the value of b to 3 significant figures.

(4)

(d) Give an interpretation of the value of a

(1)

(e) Estimate her household's daily energy consumption when the average daily temperature is 2°C

(2)

The scientist wants to use the linear regression model to predict her household's energy
consumption in the summer.

(f) Discuss the reliability of using this model to predict her household's energy consumption in the summer.

(2)

(Total for question = 14 marks)

Q17.

A farmer collected data on the annual rainfall, x cm, and the annual yield of peas, p tonnes per acre.

The data for annual rainfall was coded using $v = \frac{x - 5}{10}$ and the following statistics were found.

$$S_{vv} = 5.753 \quad S_{pv} = 1.688 \quad S_{pp} = 1.168 \quad \overline{p} = 3.22 \quad \overline{v} = 4.42$$

(a) Find the equation of the regression line of p on v in the form $p = a + bv$.

(4)

(b) Using your regression line estimate the annual yield of peas per acre when the annual rainfall is 85 cm.

(2)

(Total 6 marks)

Q18.

A researcher believes that parents with a short family name tended to give their children a long first name. A random sample of 10 children was selected and the number of letters in their family name, x, and the number of letters in their first name, y, were recorded.

The data are summarised as:

$$\sum x = 60, \quad \sum y = 61, \quad \sum y^2 = 393, \quad \sum xy = 382, \quad S_{xx} = 28$$

(a) Find S_{yy} and S_{xy}

(3)

(b) Calculate the product moment correlation coefficient, r, between x and y.

(2)

(c) State, giving a reason, whether or not these data support the researcher's belief.

(2)

The researcher decides to add a child with family name "Turner" to the sample.

(d) Using the definition $S_{xx} = \sum (x - \overline{x})^2$, state the new value of S_{xx} giving a reason for your answer.

(2)

Given that the addition of the child with family name "Turner" to the sample leads to an increase in S_{yy}

(e) use the definition $S_{xy} = \sum (x - \overline{x})(y - \overline{y})$ to determine whether or not the value

of r will increase, decrease or stay the same. Give a reason for your answer.

(2)

(Total 11 marks)

Q19.

The age, t years, and weight, w grams, of each of 10 coins were recorded. These data are summarised below.

$$\sum t^2 = 2688 \qquad \sum tw = 1760.62 \qquad \sum t = 158 \qquad \sum w = 111.75 \qquad \sum S_{ww} = 0.16$$

(a) Find S_u and S_{tw} for these data.

(3)

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between t and w.

(2)

(c) Find the equation of the regression line of w on t in the form $w = a + bt$

(4)

(d) State, with a reason, which variable is the explanatory variable.

(2)

(e) Using this model, estimate

(i) the weight of a coin which is 5 years old,
(ii) the effect of an increase of 4 years in age on the weight of a coin.

(2)

It was discovered that a coin in the original sample, which was 5 years old and weighed 20 grams, was a fake.

(f) State, without any further calculations, whether the exclusion of this coin would increase or decrease the value of the product moment correlation coefficient. Give a reason for your answer.

(2)

(Total 15 marks)

Q20.

A teacher took a random sample of 8 children from a class. For each child the teacher recorded the length of their left foot, f cm, and their height, h cm. The results are given in the table below.

f	23	26	23	22	27	24	20	21
h	135	144	134	136	140	134	130	132

(You may use $\sum f = 186$ $\sum h = 1085$ $S_{ff} = 39.5$ $S_{hh} = 139.875$ $\sum fh = 25291$)

(a) Calculate S_{fh}

(2)

(b) Find the equation of the regression line of h on f in the form $h = a + bf$. Give the value of a and the value of b correct to 3 significant figures.

(5)

(c) Use your equation to estimate the height of a child with a left foot length of 25 cm.

(2)

(d) Comment on the reliability of your estimate in (c), giving a reason for your answer.

(2)

The left foot length of the teacher is 25 cm.

(e) Give a reason why the equation in (b) should not be used to estimate the teacher's height.

(1)

(Total 12 marks)

Examiner's Report

Q1.

This proved to be a good opening question with nearly all the students making some progress. Part (a) was answered very well but in part (b), few could work out the effect of the coding on S_{ss} although more were successful with S_{st}. Some reminded us that coding had no effect on the correlation coefficient and assumed that this must also apply to these values as well, whereas others embarked on long, and usually totally incorrect, calculations having missed the significance of the instruction to "write down...".

In part (c) most tried using a correct formula and many, quite sensibly, used the variables w and t but often their values of S_{ss} and S_{st} gave an answer of $|r| > 1$ and this also meant they lost then chance of scoring in part (d). It is disappointing that many students still don't seem to realise that $|r|$ cannot be greater than 1 and use this fact to try and fix their previous errors. Those who had a valid answer for r were usually able to give a suitable comment in part (d) though some failed to emphasise that it was the magnitude of the correlation, rather than the sign, that was significant. Part (e) was often fully correct but common errors were to use $b = \frac{-49}{50}$ presumably caused by a failure to identify the response and explanatory variables correctly. The usual loss of a mark for accuracy occurred where students prematurely rounded their gradient. A few students still don't appreciate that values of a and b given as fractions are inappropriate but it was encouraging to see very few cases of students giving their final equation in terms of x and y rather than t and w as required. There were many correct responses to part (f), but some did not spot the simple link to part (e) and attempted to calculate a new regression equation usually with little success. The responses to part (g) were disappointing with many simply substituting $t = 1$ into their equation rather than giving a suitable interpretation of the gradient. Others failed to give a full enough interpretation and simply said "a decrease" whilst other were confused over the units and were happy to quote a loss of 6.5p per week.

Q2.

This proved to be a good opening question with nearly all students making some progress. Part (a) was answered successfully with most writing down a correct expression which was, of course, required for a "show that" question. Some forgot to write down the "= 5601" and lost the second mark.

In part (b), most chose the variable w but their reasons were not always precise enough. The question was phrased using the word "response" and so the examiners were looking for an explanation of this, rather than a statement such as "w responds to d". The key idea is that the number of wiggles depends on the distance of the food supply from the hive and this was all that was required. Some gave good descriptions of correlation but failed to mention the underlying causal relationship.

Part (c) was, as usual, answered very well but there were still a number of students only giving their answer to 2 significant figures and not the 3 significant figures that is standard on S1. More than 3 significant figures is fine, but answers given with fewer than 3 significant figures will lose a mark.

The calculations for part (d) were carried out confidently but a number lost the final accuracy mark by failing to carry enough figures in their calculations to achieve 3sf accuracy in the final equation and a few confused 3 significant figures with 3 decimal places. Most had a sufficiently accurate regression equation to obtain the mark in part (e) (i) and many gave a correct answer to (ii) too: some referring to interpolation and others stating clearly that the distance was in the range of the data. Common reasons for failing to score this mark were simply stating that "it" was in the range of the data without specifying that "it" is the

independent variable we are interested in.

Q3.

Part (a) was, as usual, answered very well but a number of candidates lost the final A1 because they did not round their answers to 3sf or, more worryingly, they thought that $S_{hh} = 149$ to 3sf. Most knew how to calculate r in part (b) too but few gave a full answer to part (c). Many stated that there was negative correlation (although some thought this meant that the use of a regression equation was not suitable) but few stated clearly that the use of a regression equation was suitable because there was strong correlation. Some simply said that "the points were close to a straight line" but there was no scatter diagram to support this and without a clear statement that the strong correlation suggests this the examiners could not award the mark.

Most candidates (even those who felt that a regression equation was not appropriate!) could carry out the calculations in part (d) although a sizeable minority used S_{tt} instead of S_{hh} which gave them a somewhat unrealistic gradient of − 60.3. Most found a correct gradient but often rounded their answer before calculating the intercept and the final mark was frequently lost. Full interpretations in part (e) were rare with candidates failing to mention the drop in temperature or the rise in height above sea level or give their value. The final part was answered quite well with most candidates substituting values of 500 and 1000 into their equation, only the better candidates realized that the answer was easily found from $500b$. A number of candidates seemed perfectly content with a final answer of around 30 000 °C here (due to their incorrect gradient in part (d)) and lost the final mark. Candidates should be encouraged to try and engage with the context of the questions and this can help them both in interpreting their statistical calculations and assessing the reasonableness of their answers.

Q4.

Part (a) was answered very well although some only rounded their value of b to 3 decimal places rather than 3 significant figures as required. Most knew how to start part (b) and made a correct substitution but errors often occurred when simplifying to the required form. The final part was answered well with most substituting $g = 100$ into their answer for part (b) however earlier errors meant that the correct answer was less frequently seen.

Q5.

(a) proved an accessible opening to the paper and nearly all the candidates answered this correctly. Most also knew how to find r in part (b) but a number still gave their final answer as −0.71 rather than the
3 significant figures requested. In (c) many identified that the required correlation would be positive and gave a simple argument based on the context, although a few linked the variables v and g via the third variable t. A number of candidates seemed to misread this part of the question and gave a description of the negative correlation between t and v.

Q6.

Part (a) was answered well with only a small minority using $4305 - \left(\frac{181}{8}\right)^2$. Substitution into the formula for r was carried out successfully but a number of candidates gave their final answer to only
2 significant figures instead of the standard 3 significant figures we look for on S1. Most candidates now realised that the instruction "interpret" requires a contextualised comment but there were a number of nonsensical comments such as "temperature increases as sea level decreases" which gained no credit. Most candidates knew that coding had no effect on the correlation coefficient and picked up the mark for part (e) but very few scored the mark

for part (d) with the commonest error being to divide by 1000. It appears that the effect of coding is being remembered as a fact rather than being deduced from an understanding of the structure of the formula.

Q7.

This proved to be a friendly starter for most candidates with many scoring all 5 marks in part (a) and (b). Most errors here were arithmetic such as writing S_n = 596.666... rather than 569.666... or accuracy problems in part (b) where an answer of 0.57 was often seen, rather than the 3sf accuracy that we look for. A minority still have difficulty in using the printed formulae and $S_{ll} = 327754.5 - \left(\frac{4027}{50}\right)^2$ or $\sum l \times \sum w$ instead of $\sum lx$ in S_{lw} were sometimes seen.

Part (c) caused problems for many candidates who simply wrote "positive correlation" but did not interpret this statement in the context by mentioning that longer salmon usually weigh more. Some candidates tended to "overstate" their conclusion by implying that as a salmon grows it gets longer (not strictly true in this instance as the study was of 50 different salmon not one salmon at 50 different time intervals) and others referred to a proportionate relationship such as "for every cm increase in length the salmon weighs 0.572 kg more". Whilst such indiscretions were overlooked for the single mark on this occasion, these examples should provide useful points of discussion for teachers with future cohorts of students.

Q8.

Typically candidates successfully used the correct formula in order to calculate the product moment correlation coefficient in part (a). However, a number of candidates lost the accuracy mark by only giving a rounded answer to 2 decimal places. Providing an interpretation of their value of the correlation coefficient was less straightforward. Most frequently candidates made general remarks and described the correlation as positive without relating this to the context of the question. Of those who did attempt to provide an interpretation, many failed to appreciate that it was the attendance at the matches being compared to the total number of goals scored and not the number of home matches that were played.

Part (c) was answered well overall and correct answers were often justified by accompanying statements which indicated that linear coding does not affect the product moment correlation coefficient. Some candidates, however, seemed unaware of this fact and a common mistake was to divide their original product moment correlation coefficient by 100. In addition many candidates failed to recognise the significance of them being asked to write down their answer and chose to perform a full calculation in order to obtain the product moment correlation coefficient, which sometimes led to processing errors.

Q9.

Most realised what was required in part (a) and although some went "round the houses" to solve their two simultaneous equations the correct answers were usually obtained. Part (b) was a simple application of the formula but sadly a number of students gave their answer as – 0.75 rather than an answer rounded to 3 sf. Most answered part (c) correctly with only a minority re-calculating the coefficient. Part (d) was not answered very well with many not appreciating that the correlation they had established was between distance from the station and price per square metre. Many thought they were looking at negative correlation between price and distance from the station and understandably chose *J*. Some students though did calculate the price/m^2 for each house but then some still chose house *J* as they could not interpret the negative correlation correctly but there were plenty of fully correct answers showing a thorough understanding of the situation. A few students ignored all the statistics and based their choice on other reasons such as it is "noisy near the station" and therefore the house would be cheaper.

Q10.

This question was on familiar territory and was answered well. Most candidates could carry out the calculations in part (a) successfully although a small minority seemed unfamiliar with the formulae in the formula booklet and we had some using 23070 on the numerator of r. Some lost the accuracy mark for rounding to 0.82, or even 0.8, without first stating a more accurate value.

Part (b) was usually answered correctly but a significant minority launched into some complex decoding calculations or simply left it blank. The wording "write down" and the tariff of just 1 mark should indicate that complex calculations are not required.

In part (c) most recognised that a positive correlation suggested support for the bank's claim and scored both marks. Some just stated that the correlation was strong (failing to appreciate the importance of it being positive) and a few gave sociological rather than statistical reasons.

Q11.

There seemed to be a slightly improved understanding of the calculation of standard deviation this summer and many correct responses to part (a) were seen. The usual errors still occurred though: forgetting the square root, or failing to divide the 985.88 by 8 or simply confusing S_{xx} with the variance.

The responses to part (b) confirmed that few students are aware of the simple relationship between S_{yy} and standard deviation as many attempted to find $\sum y^2$ by "unpacking" the value for δ_y with varying degrees of success. The calculation for S_{xy} in part (c) was carried out successfully by most students and many completed part (d) correctly too, usually by first finding Sxx rather than using their answer to
part (a).

Responses to part (e) were very mixed. Some simply quoted a statement about coding not affecting correlation and therefore expected no change. Others realised that the new point was breaking the trend (both values above their means conflicting with the negative correlation) but were not precise enough about the effect on the value of the correlation coefficient. The examiners were looking for a statement about the value that told us that the correlation was weaker, i.e. that the value was closer to zero:
r "increases" or r is "less negative" were not deemed to be satisfactory.

Q12.

The calculations were generally carried out very well here but the final 3 parts, requiring the students to engage with the context, were answered less well.

Despite mentioning the expectation of 3sf accuracy a number of students lost a mark here as the only evaluated answer they gave was 0.96 rather than 0.962. In question (b) most gave a comment about strong correlation or commented that the value of r was close to 1 but some misinterpreted the question and explained why linear regression was a useful tool rather than justifying its use in this situation.

In question (c) a number gave an answer of 0.74 rather than 0.740 as required but this value was allowed for the final equation in question (d) which many obtained correctly. It was rare to see a student give their equation in terms of y and x rather than m and v.

Question (e) was found to be challenging. Many confused the gradient with correlation and simply said that as the number of visitors increased so did the amount of money spent. Those who did get the idea of rate often failed to give the correct numerical values or they

had visitors and money the wrong way around e.g. "for every £1000 spent the number of visitors increased by 740" . The better answers identified that *b* represented the amount of money spent per visitor and gave simple answers such as "each visitor spends £740".

The most common error in question (f) was for students to simply substitute 2 500 000 into their equation but those who realised that 2500 should be substituted usually arrived at the correct answer and often went on to affirm the reliability of their estimate by pointing out that 2500 was within the range of the given data in question (g). Stating "it is within the range of the data" in part question (g) did not secure the marks unless the students made it clear that their "it" referred to the number of visitors and not the amount of money spent.

Q13.

Question (a) was answered very well by almost all the students but some missed the "show that" part of the demand and only scored 4 of the 5 marks here. Question (b) and question (c) were answered well and most knew how to tackle question (d) but some forgot to multiply their answer by 100 and gave an answer of £37 rather than £ 3700. A similar problem arose in question (e) with few explaining that the cost increases by £82.40 for each additional employee. Most knew the answer to question (f)(i) but question (f)(ii) was more challenging and only the more able students realised that the value of *b* would be halved.

Q14.

(a) was answered very well and only a handful of candidates did not secure the 4 marks here. Most knew how to find the equation of the regression line but sometimes candidates failed to use a sufficiently accurate value of *b* to ensure that their value of *a* was accurate to three significant figures and they therefore lost the final accuracy mark. A growing number of candidates are giving their coefficients as fractions, presumably because their calculators are set in this mode. Whilst such answers were accepted, they are not as useful as coefficients of a regression line and arguably not really appropriate in this branch of statistics.

(c) was answered very well but in (d) some candidates' responses were vague: a comment that "it is reliable because it is in the range" was not accepted because "it" does not clearly refer to the temperature. Some candidates used the technical terms of "interpolation" or "not extrapolation" correctly and these were accepted.

Q15.

Everyone could get started on this question but part (b) and drawing the line in (d) proved challenging for all but the best candidates.

Part (a) was answered well with only minor accuracy errors from some.

Few answered part (b) correctly with most trying to explain the purpose of linear regression rather than justifying its use in this instance.

Part (c) was answered very well and most knew how to use their answers in part (d) to find the gradient and intercept. A failure to work with sufficient figures on their calculators meant that many gave *a* as 0.738 which was acceptable for the first mark but not the 3sf, or better, accuracy we were looking for in the final answer. Very few gave their line as an equation of *y* on *x*.

In part (e) most plotted the mean point correctly but drawing the regression line accurately defeated many. A common error was to draw a line of best fit "by eye" that passed through the origin and a few weaker candidates simply joined up the dots.

In part (f) many correct answers were seen, mostly from substituting into the equation. Some used their graph (showing clearly on the diagram the line from $p = 16$ and from the regression line to the t axis) but this approach often lost the accuracy mark.

Q16.

There was plenty in this question for all of the students with parts (b), (c) and (e) providing a good source of marks for everyone but (a), (d) and (f) proving to be more challenging.

In part (a) the first mark was available for a sensible reason and many were prepared to have a go at giving an answer here. Better students gave 2 clear reasons for using statistical **models** whereas many started mentioning reasons for carrying out statistical **tests**. Most scored full marks in part (b) but a number mis-read 255 as 225 and lost accuracy and others lost time by calculating S_{xx}. The calculations required in part (c) were well rehearsed but a few did not use the negative value of b correctly and some failed to write down the final equation or left their values of a or b as fractions. There were many correct answers to part (d) but some gave an interpretation of the gradient not the intercept and others simply said that 28.1 was the "y intercept" without attempting to give any "contextualised" interpretation.

Part (e) was straightforward and nearly everyone scored the M mark here. Part (f) was often answered well as students realised that because the model was based on data collected in winter it would be unreliable to use it for summer temperatures as these would be different or involve extrapolation. The word "discuss" prompted some to surmise what the effect of airconditioning units would be in the summer, and this often led to contradictory statements, whilst others felt that they needed to calculate the product moment correlation coefficient and then deduced that the model was reliable.

Q17.

Part (a) was answered well by many but a number still have difficulties in determining which variable corresponds to y and which to x and then using the formulae given in the booklet. The question clearly stated that the regression line of p on v was required and this was further emphasised by giving the form as $p = a + bv$ but despite this a number calculated 1.688/1.168 for b and some used $a = 4.42 - 3.32b$. There was the usual crop of accuracy errors with candidates failing to work accurately enough to give their final coefficients to the usual 3sf (or better) accuracy. A significant number of candidates simply substituted 85 into their equation and received no marks for part (b) but a good many did appreciate the need to find the value for v and often went on to obtain an answer rounding to 4.3 as well.

Q18.

Parts (a) and (b) were answered very well with only minor slips causing a loss of marks in a few cases. In part (c) most candidates realized there was positive correlation but some went on to state that this suggested support for the researcher's belief and only the more astute explaining that the researcher should have been expecting a negative correlation and these data therefore did not offer support. Parts (d) and (e) were challenging. In part (d) many stated that S_{xx} would remain the same but they were unable to provide an adequate reason. In part (e) most thought that r would stay the same giving the text book reason that "it is not affected by coding" but a few did realize that S_{xy} would stay the same and so the increase in S_{yy} meant that r would in fact decrease.

Q19.

This question proved to be a good discriminator for the highest achieving candidates.

Part (a) and part (b) were well completed by most candidates, the biggest problem being giving r as -0.91 despite being told to give the answer to 3 significant figures.

Part (c) was generally very well answered but early rounding or lack of understanding of the difference between decimal places and significant figures led to the final equation not being stated accurately. Some candidates clearly did not know what an explanatory variable was in part (d), preferring to ignore the question and state what the dependent and independent variables were instead. A significant minority did not know what a variable was and suggested a or b.

In part (e) candidates who had the right equation had no problem getting the first answer, but a number then went on to find the answer when $t = 4$, rather than finding the change over four years. Some candidates seemed to think that they were just required to say whether the weight increased or decreased rather than find an amount, whilst others wrote a 'decrease of minus 0.1'. A few candidates believed that the coin could have increased in weight.

In part (f) some candidates just stated decrease or increase with no reason given. A number responded that it would increase and clearly understood that that the correlation would be stronger, but ignored that the result would make r closer to -1. A number of candidates said that the correlation was not affected by outliers and some thought that the removal of the fake coin constituted coding and so would have 'no effect on the product moment correlation coefficient'.

Q20.

Parts (a) and (b) were answered well by the majority of candidates. Only a small minority used $\frac{S_{fh}}{S_{hh}}$ for b and there were few cases of the incorrect sign being used for finding a. Premature rounding and a failure to write their final answer in terms of f and h to 3significant figures meant a number lost the final accuracy mark. Part (c) was answered very well but in part (d) the candidates comments were often a little confused. The question was looking for a comment that the value of the independent variable f was within the range of the data and therefore the estimate should be reliable. A number of candidates seemed to focus their comments on the value of the dependent variable h and others were just a little vague referring to "it" was within the range of the data.

There were many clear and thorough answers to part (e) that showed the candidates had a good understanding of the limitations of the equation they had calculated (namely that it was based on data for children not adults). Questions of this type are simply looking for the candidates to engage with the context and give a sensible comment.

Mark Scheme

Section 7	Correlation and linear regression part 1
qu.1	June 2017 S1 - Q1 all (d*)
qu.2	June 2016 S1 - Q1 all (e*)
qu.3	May 2013 S1 - Q1
qu.4	May 2013 S1R - Q1
qu.5	January 2013 S1 - Q1
qu.6	May 2011 S1 - Q1
qu.7	January 2011 S1 - Q1
qu.8	May 2010 S1 - Q1
qu.9	June 2015 S1 - Q2
qu.10	May 2012 S1 - Q2
qu.11	June 2016 S1 - Q3bcde*
qu.12	June 2014 S1 - Q3 all (g*)
qu.13	June 2014 S1R - Q3 all (f*)
qu.14	January 2013 S1 - Q3
qu.15	May 2012 S1 - Q3
qu.16	June 2015 S1 - Q4 all (f*)
qu.17	January 2011 S1 - Q4
qu.18	May 2013 S1R - Q5
qu.19	January 2012 S1 - Q5 all (f*)
qu.20	May 2011 S1 - Q7 all (d*)

Q1.

Question Number	Scheme	Marks
(a)	$[S_{wt}] = 784 - \frac{119\times 42}{6} =$, $\underline{-49}$	A1
	$[S_{tt}] = 2435 - \frac{119^2}{6} =$, $74.8\dot{3}$ or $74\frac{5}{6}$ or $\frac{449}{6}$ (accept awrt $\underline{74.8}$)	M1 A1 (3)
(b)	$S_{ss} = 5\times 10^7$ or **$\underline{50\ 000\ 000}$** (o.e.) $S_{st} = \underline{\mathbf{-\ 49\ 000}}$	B1 B1ft (2)
(c)	$r = \frac{"-49"}{\sqrt{50\times "74.8\dot{3}"}}$ or $\frac{"-49\ 000"}{\sqrt{"5\times 10^7"\times "74.8\dot{3}"}}$ =, $-0.80105...$ = awrt $\underline{\mathbf{-0.801}}$	M1, A1 (2)
(d)	r is close to -1 or $\lvert r\rvert$ is close to 1 or "**strong**" (o.e.) [negative] **correlation** … so "**yes**" or **does** support the belief	B1ft (1)
(e)	$b = \frac{"-49"}{"74.8\dot{3}"} = [-0.6547...]$, $a = \frac{42}{6} - b\times\frac{119}{6} = [19.9866...]$ or $a = 7 - b\times 19.8\dot{3}$	M1, M1
	So $\underline{\boldsymbol{w = 20.0 - 0.655t}}$	A1 (3)
(f)	$\underline{\boldsymbol{s = 20\ 000 - 655t}}$ or $\underline{\boldsymbol{c = 20\ 000}}$ and $\underline{\boldsymbol{d = -655}}$	B1ft B1ft (2)
(g)	**Decrease** in sales of [£] **655** (ignore any minus sign)	B1ft (1) **[14]**

	Notes
(a)	M1 for a correct expression for S_{wt} or S_{tt} (May be implied by either correct answer) 1st A1 for $[S_{wt}] = -49$ 2nd A1 for $[S_{tt}]$ = awrt 74.8 **SC** If both values correct but clearly mislabelled award M1A0A1
(b)	2nd B1ft for multiplying their S_{wt} by 1000
(c)	M1 for a correct expression using their values provided S_{tt} and S_{ss} both > 0 A1 for awrt -0.801 (Correct ans. only M1A1, -0.80 with no working M1A0)
(d)	B1ft for a correct comment that uses their value of r as support, provided $0.5 \le \lvert r\rvert \le 1$ For $\lvert r\rvert < 0.5$ comment must be "does not support", because "weak" (o.e.) correlation. **NB** "points lie close to a straight line" is B0 unless supported by mention of their value of r
(e)	1st M1 for a correct expression for b or awrt -0.66 or -0.65 Ft their answers from (a) 2nd M1 for a correct expression for a ft their value for b A1 for a correct equation in w and t only with $a = 20$ or awrt 20.0 and b = awrt -0.655 (No fractions) **If their a and b are given to more than 3 sf, accept answers in (f) to 3sf or better.**
(f)	1st B1 ft for correct c or "their 20.0" $\times 1000$ 2nd B1ft for correct d or their "-0.655" $\times 1000$ Values can be in an s, t eq'n or $c =$, $d =$ (Their a and b needn't be to 3 sf and ft their letter for t)
(g)	B1ft for stating clearly both decrease (o.e.) and [£] 655. Ft their d and allow "increase" if $d > 0$

Q2.

Question Number	Scheme	Marks
(a)	$S_{dw} = 13833 - \dfrac{"1960" \times 33.6}{8}$ or $13833 - \dfrac{65856}{8}$ (But $13833 - 8232$ is M0) $= \underline{\mathbf{5601}}$ (*)	M1 A1 cso (2)
(b)	w, since the number of wiggles depends on the distance or w depends on d	B1 (1)
(c)	$r = \dfrac{5601}{\sqrt{394600 \times 80.481}}$, $= 0.99389\ldots$ awrt $\underline{0.994}$	M1,A1 (2)
(d)	$b = \dfrac{5601}{394600}$, $= 0.014194\ldots$ (awrt 0.014) $a = \dfrac{33.6}{8} - "0.01419..." \times \dfrac{"1960"}{8} = 4.2 - "0.01419..." \times 245$ [= 0.72244..] $\underline{w = 0.722 + 0.0142d}$	M1, A1 M1 A1 (4)
(e)	(i) [0.722 + 0.0142×350 =] awrt: $\underline{5.7}$ or $\underline{5.6}$ (ii) Reliable since 350 m is in the range of the data	B1 B1 (2) **[11 marks]**

	Notes
(a)	M1 for clear attempt to find Σd and use in a correct formula. Accept $1300 < \Sigma d < 2500$ For the M1 we can condone a single slip e.g. using 1383 instead of 13833 etc A1cso for correct Σd and 5601 only. Must see the formula and so have scored M1
(b)	B1 Must select w (or wiggles) and reason based on the idea that w is dependent on d Allow w "changes according to"/ "is determined/affected by" Must mention w and d B0 for "w is measured" or "d is explanatory/indep't" or "w can't be controlled" or "w responds to d"
(c)	M1 for a correct expression (Allow ft of their incorrect S_{dw}) A1 for awrt 0.994 (Answer only 2/2) [Answer only of 0.99 scores M1A0]
(d)	1st M1 for a correct expression for b. (Allow ft of their incorrect S_{dw}) 1st A1 for awrt 0.014 No fractions. [Answer only 2/2] Can be given at final equation. [Must come from correct formula not gradient of line from e.g. (650, 9.555) to (30, 0.725)] 2nd M1 for a correct method for a. Follow through their value of b and their Σd 2nd A1 for a correct equation for w and d with a = awrt 0.722 and b = awrt0.0142 No fractions Equation in x and y is A0 Answer only 4/4
(e)	1st B1 for awrt 5.7 or awrt 5.6 2nd B1 for a reason citing 350 (m) or mentioning d is in the range of the data and stating reliable. Allow "Interpolation (or not extrapolation) therefore reliable". Saying "5.7 (or w or just "it") is in the range" is B0 "accurate" instead of "reliable" is B0 "strong correlation" (without mention of interpolation o.e.) is B0 Apply ISW if a correct comment is seen.

Q3.

Question	Scheme	Marks
(a)	$(S_{th}) = 64980 - \frac{7150 \times 110}{9} = -22408.9\ldots$ $\quad -\underline{\mathbf{22\ 400}}$	M1 A1
	$(S_{hh}) = 7171500 - \frac{7150^2}{9} = 1491222.2\ldots$ $\quad \underline{\mathbf{1\ 490\ 000}}$	A1 **(3)**
(b)	$r = \frac{-22408.9}{\sqrt{1491222 \times 371.56}} = -0.95200068\ldots$ $\quad$ awrt $-\underline{\mathbf{0.952}}$	M1A1 **(2)**
(c)	Yes as r is close to -1 (if $-1 < r < -0.5$) or Yes as r is close to 1 (if $1 > r > 0.5$) [If $-0.5 \leq r \leq 0.5$ allow "no since r is close to 0"] [If $\lvert r \rvert > 1$ award B0]	B1ft **(1)**
(d)	$b = \frac{-22408.9}{1491222.2} = -0.015027...$ $\quad$ (allow $\frac{-56}{3725}$) $\quad$ awrt -0.015	M1 A1
	$a = \frac{110}{9} - \text{"their } b\text{"} \times \frac{7150}{9} = (12.2 - -0.015 \times 794.4), = 24.1604...$ so $t = \mathbf{24.2 - 0.015}h$	M1, A1 **(4)**
(e)	0.015 is the drop in temp. (in ^{0}C), for every 1(m) increase in height above sea level.	B1 **(1)**
(f)	Change = ("24.2 – 0.015" × 500) – ("24.2 – 0.015" × 1000) or 500 × "0.015" $= \pm 7.5$ $\quad$ (awrt $\pm$ 7.5) $\quad$ (only ft a value < 100)	M1 A1ft **(2)** **(13 marks)**

	Notes
(a)	M1 for at least one correct expression (condone transcription error) 1st A1 for S_{hh} = awrt 1 490 000 or S_{th} = awrt −22 400 (Condone S_{xx} or S_{xy} =… or even S_{yy} =...) 2nd A1 for S_{th} = −22 400 and S_{hh} = 1 490 000 only. [This mark is assessing correct rounding] (Allow no labels but mis-labelling S_{th} as S_{hh} etc loses the final A1)
(b)	M1 for attempt at correct formula. Allow minor transcription errors of 2 or 3 digits. Must have their S_{hh}, S_{th} and given S_{tt} (3sf or better) in the correct places. Condone missing "−" Award M1A0 for awrt −0.95 with no expression seen. M0 for $\frac{64980}{\sqrt{7171500 \times 7.864}}$
(c)	B1ft must comment on supporting **and** state: high/strong/clear (negative or positive) correlation "points lie close to a straight line" is B0 since there is no evidence of this.
(d)	1st M1 for a correct expression for b. Follow through their S_{hh} & S_{th}. Condone missing "−" 1st A1 for awrt −0.015 or allow exact fraction from rounded values. 2nd M1 for a correct method for a. Follow through their value of b 2nd A1 for a correct equation for t and h with a = awrt 24.2 and b = awrt −0.015 No fractions
(e)	B1 Must mention h (or height) and t (or temperature) and their (1 sf) value of b in a correct comment
(f)	M1 for a correct expression seen based on their equation. Allow transcription error of 1 digit. If answer is 500× their b to 2sf and < 100 (M1A1), If answer is 500× their b to 2sf and ≥100 (M1A0)

Q4.

Question	Scheme	Marks
(a)	$b = \frac{18.35}{312.1}[= 0.058795...]$	M1
	$a = 5.8 - "0.058795..." \times 4.8$	M1
	$\underline{a = \text{awrt } 5.52}$	A1
	So $\underline{y = 5.52 + 0.0588x}$	A1 (4)
(b)	$\frac{e}{10} = "5.52" + "0.0588" \times \left(\frac{g-60}{4}\right)$	M1
	$4e = 220.71 + 0.588(g - 60)$	dM1
	$\underline{e = 46 + 0.15g}$	A1A1 (4)
(c)	$e = "46" + "0.15" \times 100$	M1
	$= \underline{61}$	A1 (2)
		[10]

	Notes
(a)	1st M1 for a correct expression for b 2nd M1 for a correct expression for a – ft their value of b 1st A1 for a = awrt 5.52 2nd A1 for a correct equation in y and x with a and b correct to awrt 3 sf
(b)	1st M1 for substitutions into their equation to get an equation in e and g. Need $y = \frac{e}{10}$ and $x = \frac{g-60}{4}$ 2nd dM1 Dep. on 1st M1 for an attempt to simplify (at least removing fractions). Allow one slip 1st A1 for an equation e = awrt $46 \pm$ ..... 2nd A1 for an equation e = ...+ awrt $0.15g$
ALT	1st M1 for use of $d = \frac{10 \times "\text{their } b"}{4}$ or sight of 0.15 used as gradient 2nd dM1 Dep. on 1st M1 for use of $\overline{e} = 10 \times "\text{their } \overline{y}"$ or sight of 58 and use of $\overline{g} = 4 \times "\text{their } \overline{x}" + 60$ or sight of 79.2 and use of these values to find c in $c = \overline{e} - d\overline{g}$
(c)	M1 for substituting $g = 100$ into their new equation (or $x = 10$ and then attempting to $\times$ ans.by 10) A1 for awrt 61

Q5.

Question Number	Scheme	Marks
(a)	$(S_{tt}) = 8702 - \frac{258^2}{10}$ or $(S_{gt}) = 1550.2 - \frac{258 \times 63.6}{10}$	M1
	$(S_{tt} =) 2045.6$, $(S_{gt} =) -90.68$ awrt (2046), awrt – 90.7	A1, A1
		(3)
(b)	$r = \frac{-90.68}{\sqrt{2045.6 \times 7.864}} = -0.714956...$ awrt -0.715	M1 A1
		(2)
(c)	Positive e.g. high v corresponds to low t and low t corresponds to high g so expect high v to corresponds to high g or expect more revision to result in a better grade	B1 B1 (2)
		7

Notes

(a) M1 for at least one correct expression

1st A1 for S_{tt} = awrt 2046 (Condone S_{xx} =… or even S_{yy} =…)

2nd A1 for S_{gt} = awrt -90.7 (Condone S_{xy} =…)

(b) M1 for attempt at correct formula.

Must have their S_{tt}, S_{gt} and given S_{gg} in the correct places. Condone missing "–"

Award M1A0 for awrt -0.71 with no expression seen

M0 for $\frac{1550.2}{\sqrt{8702 \times 7.864}}$

Correct answer only is 2/2

(c) 1st B1 for saying "positive". Ignore mention of skew.

2nd B1 for suitable reason that mentions at least v and g and supports positive correlation.

e.g. "the less revision done the lower the grade" is B1

"should do better with more revision" is B0 since does not mention grades

"both coefficients are similar" or two sketches of negative correlation with labelled axes is B1 since v, t and g are implied

Allow use of letters v and g

Allow equivalent terms e.g. "study" instead of "revision" or "score" instead of "grade"

Q6.

Question Number	Scheme	Marks
(a)	$S_{yy} = 4305 - \frac{181^2}{8}$ $= \underline{209.875}$ **(awrt 210)**	M1 A1 (2)
(b)	$r = \frac{(-)23726.25}{\sqrt{3535237.5 \times "209.875"}}$ $= \underline{-0.87104\ldots}$ **(awrt −0.871)**	M1 A1 (2)
(c)	Higher towns have lower temperature or temp. decreases as height increases	B1 (1)
(d)	$S_{hh} = 3.5352375$ **(awrt 3.54)** (condone 3.53)	B1 (1)
(e)	$r = \underline{-0.87104\ldots}$ **(awrt −0.871)**	B1ft (1) **(7 marks)**

Notes

(a) M1 for a correct expression. Allow one slip e.g. 4350 for 4305

(b) M1 for a correct expression for r, follow through their answer to (a). Condone no "_ "

Allow M1 for $\pm$ 0.87 with no working. (−0.871 is M1A1)

(c) B1 Must mention <u>temperature</u> (o.e.) and <u>height</u> (above sea level) and interpret the relationship between them. Must be a correct <u>and sensible</u> comment.
e.g. "As temperature increases the height <u>of the sea</u> decreases" is B0. BUT simply stating "As temperature increases the height decreases" is B1 although "As height increases the temperature decreases" would be better. Treat mention of 0.87... as ISW

"strong negative correlation between height and temp" is B0 (no interpretation)
" as x increases y decreases" is B0 (no mention of height and temperature)

(d) B1 accept awrt 3.54 and condone 3.53 (i.e truncation)

(e) B1ft for awrt −0.871
or ft their final answer to part (b) to the same accuracy (or 3 sf) provided $-1 < r < 1$

Answer to part (e) must be a number "it's the same" is B0

Q7.

Question Number	Scheme	Marks
(a)	$S_{ll} = 327754.5 - \frac{4027^2}{50} = 3419.92$ $S_{lw} = 29330.5 - \frac{357.1 \times 4027}{50} = 569.666$	M1 A1 A1 (3)
(b)	$r = \frac{569.666}{\sqrt{3419.92 \times 289.6}} = 0.572$ awrt 0.572 or 0.573	M1 A1 (2)
(c)	As the length of the salmon increases the weight increases	B1ft (1) [6]

Notes

(a)

M1 for at least one correct expression

1st A1 for S_{ll} = awrt 3420 (Condone S_{xx} =… or even S_{yy} =…)

2nd A1 for S_{lw} = awrt 570 (Condone S_{xy} =…)

(b)

M1 for attempt at correct formula.

Must have their S_{ll}, S_{lw} and given S_{ww} in the correct places

If S_{ll}, S_{lw} are correct and an answer of awrt 0.57 is seen then award M1A0

M0 for $\frac{29330.5}{\sqrt{327754.5 \times 289.6}}$

(c)

B1ft for a comment mentioning "length" and "weight", not just *l* and *w*, and the idea of longer salmon weighing more.

e.g. "positive correlation between weight and length" is B0 since the idea of positive correlation is not explained.

Allow "larger" instead of "heavier" or "longer"

Ignore any spurious values mentioned such as 0.572

If their r is negative (but must be $r > -1$) ft an appropriate comment.

Condone $r > 1$ if comment is correct.

If $|r| < 0.4$ allow a comment of no or little relationship between weight and length but for $0 < r < 0.4$ the printed answer is still acceptable too.

Treat mention of "skewness" as ISW if a correct interpretation is given

Q8.

Question Number	Scheme	Marks
(a)	$r = \dfrac{8825}{\sqrt{1022500 \times 130.9}}$, = awrt **0.763**	M1 A1 (2)
(b)	Teams with high attendance scored more goals (oe, statement in context)	B1 (1)
(c)	0.76(3)	B1ft (1)
		Total 4
(a)	M1 for a correct expression, square root required Correct answer award 2/2	
(b)	Context required (attendance and goals). Condone causality. B0 for 'strong positive correlation between attendance and goals' on its own oe	
(c)	Value required. Must be a correlation coefficient between -1 and +1 inclusive. B1ft for 0.76 or better or same answer as their value from part (a) to at least 2 d.p.	

Q9.

Question	Scheme	Marks
(a)	$\frac{1840-a}{b}=4.0$ $\quad$ $\frac{1848-a}{b}=4.8$ $a=\underline{\mathbf{1800}}$ $\quad$ $b=\underline{\mathbf{10}}$	M1 A1 (2)
(b)	$r=\frac{-2.17}{\sqrt{1.02\times 8.22}}=-0.749417343....$ $\quad$ awrt $-\underline{\mathbf{0.749}}$	M1A1 (2)
(c)	-0.749	B1ft (1)
(d)	House H: 156400/85 = [£1840/m^2 or $q=4$] House J: 172900/95 = [£1820/m^2 or $q=2$] Since ($r=-0.749$,) there is negative correlation. <u>or</u> The higher the price (per square metre), the lower the distance from the train station. Therefore…..House H is likely to be closer.	M1 dM1 A1 (3) (8 marks)

	Notes
(a)	M1 for setting up two suitable equations which could lead to a and b (may be implied by one correct answer) A1 for $a=1800$ <u>and</u> $b=10$ ($a=10$ and $b=1800$ is A0) Correct answer only is 2/2
(b)	M1 for a correct expression (condone missing $-$) A1 for awrt -0.749 (-0.75 <u>or</u> awrt 0.749 with no working scores M1 A0).
(c)	B1ft for -0.749 or ft their answer to (b) to at least 2sf. Must be in the range $-1<'(b)'<1$
(d)	M1 for calculating price/square metre for <u>both</u> H and J. Can be implied by sight of 1840 and 1820 (so OK if not labelled or mis-labelled) These may be seen in the table in the question. Allow comment like "H is £20/square metre more than J" dM1 dependent on 1st M1 for a statement that correlation is negative <u>or</u> a contextualised interpretation of the negative correlation. $r>0$ If $r>0$ allow equivalent statements about positive correlation A1 (dependent on both Ms) for House H is likely to be closer (No ft if $r>0$)

Q10.

Question	Scheme	Marks
(a)	$\left[S_{xy}=\right] 23070-\frac{477\times480}{12}\quad[=3990]$	B1
	$r=\frac{"3990"}{\sqrt{5606.25\times4244}}$	M1
	$=0.81799\ldots$ awrt **0.818**	A1 (3)
(b)	**0.818**	B1ft (1)
(c)	Positive correlation or value of r is close to 1 or value of $r>0$ (NOT "high/ strong correlation")	B1
	So there is support for the bank's claim or "increase in unemployment is accompanied by increase in house repossessions"	B1 (2)
		[6]

Notes

(a) **Marks for part (a) must be seen in (a), do not award if only seen in (b)**

B1 for a correct expression for S_{xy}

M1 for correct attempt at r f.t. their 3990 but $\frac{23070}{\sqrt{5606.25\times4244}}$ is M0

A1 for awrt 0.818 If an answer of 0.82 only is seen then B1M1A0 can be given

(b) B1ft for awrt 0.818 or f.t. their answer to part (a) for $|r|<1$. Allow 2sf or 1sf follow through
Answer in (b) must be correct or match one of their answers in (a). Must be a number.

(c) 1st B1 for a reason of positive correlation (allow even if $r>1$)
"positive skew" or "positive gradient" is B0 but 2nd B1 is still possible
2nd B1 for a comment that suggest this supports the claim.
Marks in (c) are independent but first B1 requires some idea of positive correlation

(c) SC If $|r|<0.2$ allow this alternative to the mark scheme:
1st B1 for saying there is no or little correlation
2nd B1 for a comment that says this does not support the bank's claim

Q11.

Question Number	Scheme	Marks
(a)	$[\sigma_x^2 =] \frac{985.88}{8} - \left(\frac{86.8}{8}\right)^2 = \frac{985.88}{8} - 10.85^2$	M1
	$\sigma_x = \sqrt{\frac{985.88}{8} - \left(\frac{86.8}{8}\right)^2} = \sqrt{123.235 - 117.7225} = \sqrt{5.5125}$ or $\sqrt{\frac{44.1}{8}}$	A1
	$= 2.3478\ldots =$ awrt **2.35**	A1
		(3)
(b)	$S_{yy} = 8 \times \sigma_y^2 = 716$ (3 sf) but may see	
	$1136.584 - \frac{58^2}{8}$ or $27628(.084168) - \frac{464^2}{8}$ or 716.08... (= 716 to 3 sf) (*)	B1cso
		(1)
(c)	$S_{xy} = 4900.5 - 58 \times 86.8$ or $4900.5 - \frac{86.8 \times 464}{8}$	M1
	$= -133.9$ (Allow -134)	A1
		(2)
(d)	$r = \frac{-133.9/8}{\sigma_x \times \sigma_y}$ or $\frac{-133.9}{\sqrt{44.1 \times 716}}$	M1
	= awrt -0.753 or -0.754	A1
		(2)
(e)	$r < 0$ means high sunshine and low rain; this is high sunshine high rain	B1
	[this is not in keeping with the trend so] r is closer to 0 or $\lvert r \rvert$ decreases	B1
		(2)
		[10 marks]

	Notes	
(a)	M1 for a correct expr' for st. dev or variance (ignore label)[may be implied by 2.35 or 5.5125]	
	1st A1 for a correct expression for st. dev (must have square root) can ignore label	
	2nd A1 for awrt 2.35 (allow $s = 2.5099\ldots$ or awrt 2.51). If they have $\sigma^2 = 2.35$ score A0 but condone no label	
(b)	B1cso for a correct expression or sight of at least 716.08… (NB limits: 716.00~716.16)	
	Do not allow verification. Beware circular arguments: $716 \rightarrow \Sigma y^2 \rightarrow$ expr' $\rightarrow 716$	
(c)	M1 for a correct expression for S_{xy} (NB $\Sigma y = 464$)	
	A1 for -133.9 or awrt -134 [No fractions] (Answer only 2/2)	
(d)	M1 for a correct expression for r (ft their values for S_{xy} and σ_x or S_{xx})[Allow ft of S_{yy}]	
	A1 for awrt -0.753 or -0.754 (Answer only 2/2)	
(e)	**If they do not have an answer to (d) or their value of r is > 0 or $\lvert r \rvert > 1$ score B0B0 here**	
	1st B1 for a suitable reason contradicting $r < 0$ e.g. new value is not in keeping with trend or both $14 > \bar{x}$ and $70 > \bar{y}$ or saying both above average. Allow for $-0.48 <$ new $r < -0.47$	
	2nd B1 for a correct statement about r getting closer to zero e.g. $\lvert r \rvert$ decreases	
	A comment that r decreases or r is smaller or r is "less negative" is B0	
	"r increases" is B0 unless they also say that it gets closer to 0	

Q12.

Question Number	Scheme	Marks
(a)	$r = \dfrac{31512.5}{\sqrt{42587.5 \times 25187.5}} = 0.962$ awrt **0.962**	M1 A1 **(2)**
(b)	r is close to 1 or a **strong correlation**. ["points are close to a straight line" isB0] [Just "positive" correlation is B0] [Use of "relationship" or "skew" not "correlation" is B0]	B1 **(1)**
(c)	$b = \dfrac{31512.5}{42587.5} = 0.739947... = 0.740$ (3 dp) **0.740** (only)	M1 A1cao **(2)**
(d)	$a = 1326.25 - (0.7399... \times 2423.75)$ [$= -467.2$ or awrt -467]	M1
	So $m = -467 + 0.74v$	A1 **(2)**
(e)	b is the money (spent) per visitor. (i.e. definition of a rate in words.)[ignore values]	B1
	So each 1000 visitors generates an extra £0.74 million or each visitor spends £740 oe	B1ft **(2)**
(f)	$m = -467 + 0.74 \times 2500$	M1
	$m = 1383$ (£ million) awrt **1380**	A1 **(2)**
(g)	As 2500 is within the range of the data set or it involves interpolation.	B1
	The value of money spent is reliable	dB1 **(2)**
		Total 13

	Notes
(a)	M1 for a correct expression for r. Ans only of 0.96 or awrt 0.96 is M1A0 Ans only of 0.962 or awrt 0.962 is M1A1. Do not allow fractions for A1
(b)	B1 for comment implying strong correlation. (e.g. big/high/clear etc) B0 if $\lvert r \rvert > 1$
(c)	M1 for a correct expression for b (may be implied by 0.74 or better in regression equation)
	A1 A1 for 0.740 only in (c) or $b = 0.740$ seen elsewhere (M1A0 for $\dfrac{2521}{3407}$ or awrt 0.74 here)
(d)	M1 for 1326.25 − ('their b' × 2423.75) Condone fractions or awrt 1330 for $\overline{m}$ and awrt 2420 for $\overline{v}$
	A1 for a correct equation in m and v with a = awrt − 467 and b = awrt 0.74 Condone $\frac{2521}{3407}$ for b and $\frac{-1591740}{3407}$ for a. [Equation in y and x is A0]
(e)	1st B1 for a correct definition of the rate in words. Must state or imply "money per visitor" Allow alternative words or symbols e.g. £ or "pounds" for money, "people" for visitors etc
	2nd B1ft for a correct numerical rate (ft their value of b)
	e.g. "each visitor spends £740" is B1B1, "b is the extra money spent per visitor" is B1B0 [no values]
	"b is increase of £0.74 million in m as v increases by 1000" is B0B1[£ for money but no "visitors"] "increase in m as v increases" is B0B0 [Idea of rate but letters not words and no numerical value of rate]
(f)	M1 sub. $v = 2500$ into their equation. Simply substituting 2 500 000 is M0 (unless adjusted eqn)
	A1 awrt 1380 units (£ and million not required)
(g)	1st B1 for 2500 or 2 500 000 or visitors or v is in range. "it" is B0 unless v clearly implied
	2nd dB1 for stating it is reliable. Dependent on previous B mark being awarded
	"both v and m in range" or "1380 in range" is B0 but use ISW so "interpolation since both in range" scores B1 for the "interpolation". "Not extrapolation" counts as "interpolation"

Q13.

Question Number	Scheme	Marks
(a)	$\sum x = 88$	B1
	$S_{pp} = 11422 - \frac{299.2^2}{8} = [231.92]$ (*)	B1cso
	$S_{xx} = 1160 - \frac{'88'^2}{8} = 192$	M1 A1
	$S_{xp} = 3449.5 - \frac{'88' \times 299.2}{8} = 158.3$ **awrt** 158	A1
		(5)
(b)	$r = \left[\frac{S_{xp}}{\sqrt{S_{xx} S_{pp}}} = \right] \frac{'158.3'}{\sqrt{'192' \times 231.92}}$	M1
	$r = 0.7501726031\ldots$ **awrt** 0.750	A1
		(2)
(c)	$b = \left[\frac{S_{xp}}{S_{xx}} = \right] \frac{'158.3'}{'192'} = 0.824(479166....)$ (*)	M1 A1cso
	$a = \overline{p} - b\overline{x} = \frac{299.2}{8} - 0.824... \times \frac{"88"}{8} = 28.330729...$ **awrt** 28.3	M1 A1
		(4)
(d)	$p = 28.3.... + 0.824.... \times 10 = 36.57552....$ **awrt** £3700	M1 A1 **(2)**
(e)	Goes up £82.40	B1
		(1)
(f)	(i) $r = 0.750$	B1ft
	(ii) $b = 0.412$	B1 **(2)**
		[Total 16]

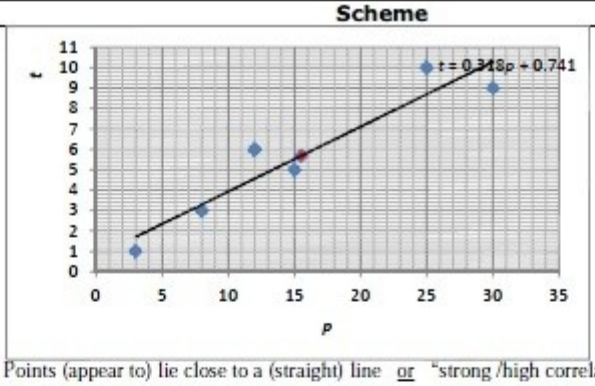
	Notes
(a)	1st B1 for $\sum x = 88$ seen. May be in a correct formula or implied by 192 or 158.3
	2nd B1cso for a correct expression for S_{pp}
	M1 for a correct expression for S_{xx} or S_{xp} (ft their Σx). If we don't see an explicit $\Sigma x = k$ but consistent use of k instead of 88 in S_{xp} and S_{xx} then award M1
	1st A1 for $S_{xx} = 192$ 2nd A1 for S_{xp} = awrt 158
(b)	M1 for correct expression for r ft their 192 and 158.3 May be implied by $r = 0.75$
	A1 for awrt 0.750 Allow A1 for $r = 0.75$ if a correct expr' is seen (since 3rd sf is 0)
(c)	1st M1 for a correct expression for b using their values NB. use of 158 gives 0.8229
	1st A1 cso for b = awrt 0.824
SC	If there is no expression but 0.8244...or better is seen award 1 mark as M0A1
	2nd M1 for a correct expression for a ft their Σx
(d)	2nd A1 for a = awrt 28.3
	M1 for substituting $x = 10$ into their equation
(e)	A1 for awrt £3700 (£ 36.58 or £36.58 (hundreds) is A0)
(f)(i)	B1 for goes up £82.40 (for each additional employee) (£0.824 hundreds is B0)
(ii)	B1ft for r = their answer to (b) . Allow recalculation. Condone $\lvert r \rvert > 1$
	B1 for 0.412 only

Q14.

Question Number	Scheme	Marks
(a)	$\sum t = 140$ (or $\bar{t} = 17.5$) and $\sum m = 32$ (or $\bar{m} = 4$)	B1 B1
	$(S_{tm}) = 469.5 - \frac{"140" \times "32"}{8}$	M1
	$(S_{tm} =) -90.5$	A1cso
		(4)
(b)	$b = \frac{S_{tm}}{S_{tt}} = \frac{-90.5}{354}$	M1
	$b = -0.255649...$ (allow $\frac{181}{708}$) -0.25 or awrt -0.26	A1
	$a = \frac{"32"}{8} - b \times \frac{"140"}{8}$	M1
	So equation of the line is $\underline{\mathbf{m = 8.47 - 0.256t}}$ (allow $m = \frac{11999}{1416} - \frac{181}{708}t$)	A1
		(4)
(c)	$(8.47 - 0.256 \times 10 =) 5.9...$ awrt $\underline{5.9}$	B1 (1)
(d)	Should be reliable since 10 is in the range (of the data)	B1 (1)
		10

	Notes
(a)	1st B1 for 140 seen in correct context or correctly labelled 2nd B1 for 32 seen in correct context or correctly labelled. (allow a fully correct expression – not "...+...+...") 4480 used correctly is B1B1 M1 for attempting a correct expression. Follow through their 140 and their 32 You may see attempt at $\sum(t-\bar{t})(m-\bar{m})$. This must have all the products seen. A1cso requires a correct expression seen and no incorrect working leading to -90.5
(b)	1st M1 for a correct expression for b. Follow through their S_{tm}. Condone missing "–" 1st A1 for awrt -0.26 or condone -0.25 2nd M1 for a correct method for a. Follow through their sums from part (a) and their value of b 2nd A1 for a correct equation for m and t with a = awrt 8.47 and b = awrt -0.256 Must be an equation in m and t, use of x or y scores A0 here.
(c)	B1 for awrt 5.9 Accept 6 if the correct expression (awrt $8.47 - 10 \times$ awrt 0.256) is seen
(d)	B1 for suggesting it is reliable and mentioning 10 within the range (of the data.) or suggesting it is reliable since interpolating or not extrapolating NB "it is reliable since it is in the range" is B0 since "it" is not explicit enough Condone extra non-relevant comments but penalise contradictory comments. e.g. "near the extreme so not reliable but not extrapolated so reliable" is B0 since contradicts "reliable since 10 is within the range (of temps) and 5.9 within range of times" is B1 since irrelevant

Q15.

Question	Scheme	Marks
(a)		Use overlay B1 B1 (2)
(b)	Points (appear to) lie close to a (straight) line or "strong /high correlation"	B1 (1)
(c)	$\sum p = 93$ and $\sum t = 34$ (may be seen in table)	M1
	$S_{pt} = 694 - \frac{"93" \times "34"}{6} = [167]$ or $S_{pp} = 1967 - \frac{"93"^2}{6} = [525.5]$	M1
	$S_{pt} = 167$; S_{pp} = awrt 526	A1; A1 (4)
(d)	$b = \left[\frac{S_{pt}}{S_{pp}} =\right] \frac{"167"}{"525.5"} = [0.31779...]$ (**check their answer if expression not seen**)	B1ft
	$a = \frac{"34"}{6} - "0.31779..." \times \frac{"93"}{6} = 5.666... - 0.31779... \times 15.5 =$, 0.74088… awrt 0.74	M1, A1
	$t = 0.741 + 0.318p$ (Accept $a = \frac{2336}{3153}$ and $b = \frac{334}{1051}$ in their equation)	A1 (4)
(e)	$(\overline{p}, \overline{t})$ = (15.5, 5.7) plotted on the graph (not wholly outside the circle)	B1
	Correct line plotted as per overlay. For $p = 5$; $2 < t < 3$ and for $p = 30$; $10 < t < 11$ Their line must stretch roughly as far as the points and go through the $(\overline{p}, \overline{t})$ circle	B1 (2)
(f)	$t = "0.741" + "0.318" \times 16$	M1
	= 5.825... **awrt 5.8**	A1 (2)
		[15]

	Notes
(a)	B2 for all 6 data points plotted correctly. B1 for any 5 correct. Points not wholly outside the circles.
(c)	1st M1 for attempting $\sum p$ and $\sum t$. Allow $80 < \sum p < 100$ and $30 < \sum t < 40$ 2nd M1 for one correct expression for S_{pt} or S_{pp}, f.t. their $\sum p$ and $\sum t$. 1st A1 for S_{pt} 2nd for S_{pp}
(d)	B1ft for correct expression for the gradient, f.t. their 167 and 525.5 from (c) M1 for correct use of $a = \overline{t} - b\overline{p}$ f.t. their values. Condone 5.6 for $\overline{t}$ 1st A1 for awrt 0.74 NB use of 526 gives 0.745566… and gets A0 2nd A1 for a correct equation for t in terms of p with a and b awrt 3sf An equn in y or x is A0
(f)	M1 for clear use of their line (equation or on graph) and $p = 16$ to estimate t. This may be an expression or lines marked on the diagram A1 for awrt 5.8, even if their line is not fully correct. Accept "$t > 5.8$" (oe). Answer only 2/2

Q16.

Question	Scheme	Marks
(a)	To simplify (or represent) a real world problem (o.e.) To improve understanding (o.e.) To analyse a real world problem or can change variables/replicate easily (oe) To make predictions or find estimates (o.e.)	B1g B1h (2)
(b)	$\sum x = 12$ $S_{xy} = 283.8 - \frac{'12' \times 255}{10}$, $= -\underline{22.2}$	B1 M1,A1cao (3)
(c)	$b = \frac{'-22.2'}{10.36} =, -2.142857....$ (A1 for awrt -2.1) $[a = \bar{y} - b\bar{x} \Rightarrow]\ a = \frac{255}{10} - 'b' \times \frac{"12"}{10} = 28.07143$ $y = 28.1 - 2.14x$ [Condone: $y = 28.1 + -2.14x$]	M1A1 M1 A1 (4)
(d)	(28.1 kWh) of energy are used when the temperature is 0[°C]	B1 (1)
(e)	$y = 28.1 - 2.14(2) =$ awrt $\underline{23.8}$	M1 A1 (2)
(f)	The regression model is based on temperatures from the winter, so not reliable in the summer. Stating it is reliable (whatever the reason) is B0B0	B1 dB1 (2) (14 marks)

	Notes	
(a)	Make sure reasons refer to **models** and not **tests** 1st B1g (be fairly generous) for a sensible reason not using "quick", "cheap" or "describe" 2nd B1h (be slightly harder) for two convincing reasons (both based on the list above) Use professional judgement and mark as B0B0 or B1B0 or B1B1 do **not** use B0B1	
(b)	B1 for $\sum x = 12$ (May be by the table) (Can be implied by 3060 seen or the next line) M1 for attempt at correct formula (ft their $\sum x$ where $10 < \sum x < 14$) A1 for -22.2 only	
(c)	M1 for a correct expression for b (ft their $S_{xy} \neq 283.8$) A1 for awrt -2.1 (allow $-15/7$) M1 for a correct expression for a and ft their 12 (allow use of a letter b) A1 for $y = 28.1 - 2.14x$ (awrt 28.1 and awrt -2.14) Must be y and x and no fractions	
(d)	B1 for a contextualised interpretation e.g. the amount of <u>energy</u> used when <u>temperature</u> is <u>0</u>[°C] or [28.1] <u>kWh</u> used when <u>temp. is 0</u>[°C] [Can ft their 28.1]Need <u>temp</u> or °<u>sign</u> [B0 for "value of y when $x = 0$" since no context in words]	
(e)	M1 for substituting $x = 2$ into their equation	
(f)	B1 for reasoning to suggest that temperatures are different in summer or the model was based only on data from the winter. Allow mention of <u>extrapolation</u> (o.e.) dB1 so not reliable.	

Q17.

Question Number	Scheme	Marks
(a)	$b = \frac{1.688}{5.753} = 0.293$ $a = 3.22 - 4.42 \times 0.293 = 1.9231...$ $p = 1.92 + 0.293v$	M1A1 M1 A1 (4)
(b)	$v = \frac{85-5}{10} = 8$ $p = 1.92 + 0.293 \times 8 = 4.3$ **(awrt 4.3)**	M1 A1 (2) **[6]**
	Notes	
(a)	**Can ignore (a) and (b) labels here** 1st M1 for a correct expression for b. $\frac{1.688}{1.168}$ is M0 1st A1 for awrt 0.29 2nd M1 for use of $a = \bar{p} - b\bar{v}$ follow through their value of b(or even just the letter b) 2nd A1 for a complete equation with a = awrt 1.92 and b = awrt 0.293 y or $p = 1.92 + 0.293x$ is A0 Correct answer with no working is 4/4	
(b)	M1 for an attempt to find the value of v when $x = 85$ (at least 2 correct terms in $\pm\frac{85-5}{10}$) or for an attempt to find an equation for p in terms of x and using $x = 85$ Attempt at equation of p in x requires $p = 1.92 + 0.293\frac{(x-5)}{10}$ A1 for awrt 4.3 (award when first seen and apply ISW) N.B. $p = 1.92 + 0.293 \times 85$ (o.e.) is M0A0	

Q18.

Question	Scheme	Marks
(a)	$S_{yy} = 393 - \frac{61^2}{10} = \underline{\mathbf{20.9}}$	M1A1
	$S_{xy} = 382 - \frac{61\times 60}{10} = \underline{\mathbf{16}}$	A1 (3)
(b)	$[r =] \frac{"16"}{\sqrt{"20.9"\times 28}}$	M1
	$= 0.66140\ldots$ **awrt 0.661**	A1 (2)
(c)	Researcher's belief suggests negative correlation, data suggests positive correlation	B1
	So data does not support researcher's belief	dB1 (2)
(d)	New x equals $\bar{x} = 6$	B1
	Since $S_{xx} = \sum(x-\bar{x})^2$ the value of S_{xx} is the same = 28	dB1 (2)
(e)	$S_{xy} = \sum(x-\bar{x})(y-\bar{y}) = \sum(x-\bar{x})y$ so the new term will be zero (since mean = x) and since S_{yy} increases	B1
	So r will decrease	dB1 (2)
		[11]

	Notes
(a)	M1 for a correct expression for S_{yy} or S_{xy} 1st A1 for $S_{yy} = 20.9$ 2nd A1 for $S_{xy} = 16$
(b)	M1 for a correct expression for r – ft their 20.9 (provided it is > 0) and their 16. Use of 382 for 16 or 393 for 20.9 is M0 A1 for awrt 0.661
(c)	1st B1 for a suitable reason contrasting belief with data. They must state the sign (positive or negative) of the correlation of data or the belief and imply the other is opposite 2nd dB1 Dependent on a correct reason for saying it does not support the claim e.g. State "does not support the belief because data has positive correlation" scores B1B1 BUT State "does support the belief because data has positive correlation" scores B0B0
(d)	1st B1 for clearly stating that new value of x = (6 =) mean 2nd dB1 Dep. on 1st B1 for a reason that shows S_{xx} is unchanged e.g. extra term is 0 so S_{xx} is the same
ALT	1st B1 for seeing $\sum x = 66$ and new $\sum x^2 = 424$ (or $388 + 6^2$) and attempt at S_{xx} 2nd B1 for showing $S_{xx} = 28$ with $n = 11$ and no incorrect working seen and a final comment
(e)	1st B1 for a clear reason that mentions S_{xy} is the same and the increase in S_{yy} Saying that r increases or stays the same is B0B0 2nd dB1 Dependent on 1st B1 for saying r will decrease.

Q19.

Question Number	Scheme		Marks
(a)	$S_{tt} = 2688 - \frac{158^2}{10} = 191.6$	awrt 192	M1 A1
	$S_{tw} = 1760.62 - \frac{158 \times 111.75}{10} = -5.03$	awrt -5.03	A1 (3)
(b)	$r = \frac{-5.03}{\sqrt{191.6 \times 0.16}} = -0.908469\ldots$	awrt -0.908(5)	M1A1 (2)
(c)	$b = \frac{-5.03}{191.6} = -0.0263$	awrt -0.026	M1 A1
	$a = 11.175 + 0.0263 \times 15.8$ $= 11.59$		M1
	$w = 11.6 - 0.0263t$		A1 (4)
(d)	The explanatory variable is the age of each coin. This is because the age is set and the weight varies.		B1 B1 (2)
(e) (i)	awrt 11.5		B1
(ii)	Decrease(in weight of coin of 0.1052 g) = 0.1 or -0.1 or increase of -0.1 awrt(-0.1)		B1 (2)
(f)	Decrease; removing the fake will result in a better linear fit so r will be closer to -1		B1;B1 (2) **Total 15**
NOTES			
(a)	M1 for correct attempt at either method, A1 awrt 192 A1 awrt -5.03		
(b)	M1 for correct attempt at use of formula, square root required. A1 awrt -0.908(5)		
(c)	M1 require 'their -5.03' as numerator and /their 191.6' as denominator. A1 awrt -0.026 M1 for use of correct formula with b or 'their b'; require $--$ or $+$ and values in the correct place. A1 for equation as written with values awrt 3 sf. with w and t. Accept fractional answers that are accurate to 3sf when evaluated as decimals		
(d)	B1 for 'Age' or t or 'years' B1 for 'you use age / t to predict w' or 'you can control t/ age' or 'weight depends on age' or similar		
(e)	B1 awrt 11.5 B1 awrt -0.1 but 'decrease of -0.1' is B0.		
(f)	B1 for Decrease only but 'mod r increases' explicitly stated in words or symbols award B1. B1 accept 'stronger correlation' or 'increase in correlation' or 'better linear fit' or ' r closer to -1' or 'points are closer to a straight line' or 'point is an outlier' or equivalent		
Special Case 1	Attempt to calculate S_{tw}		
	$\sum tw = 1669.62, \sum t = 153, \sum w = 91.75$ or $S_{tw} = 1660.62 - \frac{153 \times 91.75}{9}$ or awrt 101		
	or $S_{tw} > 0$ with some calculation		B1
	"Increase"		B1 (2)
Special Case 2	Attempt to calculate S_{ww}		
	$\sum w^2 = 1248.96625 - 400 = 848.96625$ or awrt 849 or $S_{ww} = 848.96625 - \frac{91.75^2}{9}$		
	or awrt -86.4 or $S_{ww} < 0$		B2 (2)
Special Case 3	Argument based on standard deviation.		
	e.g. $\sigma_w \approx 0.126$ and $\overline{w} = 11.175$ so fake coin is over 69 sds away from the mean		B1
	'(very) unlikely' or 'impossible'		B1 (2)

Q20.

Question Number	Scheme	Marks
(a)	$\left(S_{fh}=\right)25291-\frac{186\times1085}{8}$	M1
	$=\underline{64.75}$ (accept 64.8)	A1
		(2)
(b)	$b=\frac{"64.75"}{39.5}$, $=\underline{1.6392....}$ (awrt 1.6)	M1, A1
	$a=\frac{1085}{8}-b\times\frac{186}{8}$, $=\underline{97.512...}$ (awrt 97.5)	M1, A1
	$\underline{h=97.5+1.64f}$	A1ft (dep on M1M1)
		(5)
(c)	$h=97.5+1.64\times25$, $=\underline{138\sim139}$ (final answer in [138, 139])	M1, A1
		(2)
(d)	Should be reliable, since 25 cm(or f or footlength) is within the range of the data	B1, B1
		(2)
(e)	Line is for children – a different equation would apply to adults or Children are still growing, height will increase more than foot length	B1
		(1)
		12

	Notes
	[NB $r = 0.871$ so do not confuse this with question 1]
(a)	M1 for attempting a correct expression [allow a copying slip e.g. 25921]
(b)	1st M1 for a correct expression for b, ft their part (a) but not $S_{fh} = 25291$ 1st A1 for awrt 1.6 2nd M1 for use of $a=\overline{h}-b\times\overline{f}$, ft their value for b. Must use $\overline{h}$ and $\overline{f}$ not values from table. 2nd A1 for awrt 97.5 [NB $a = 135 - 1.63\times23 = 97.51$ but M0A0 since not using $\overline{h}$ and $\overline{f}$] 3rd A1ft for an equation for h and f with their coefficients to 3sf. **Dependent on both Ms** Must be 3sf not awrt. Give this mark if seen in (c). Equation must be in h and f not y and x.
(c)	M1 for using their equation and $f = 25$ to find h A1 for their final answer in [138, 139]. Can give if they have 137.7... but round to 138
(d)	1st B1 for suggesting it is reliable 2nd B1 for mentioning that 25 cm is within range of data. "interpolation"or"not extrapol'B1 Use of "it" or a comment that height is in range is B0 but apply ISW
(e)	B1 for some comment that states a difference between children and teachers(adults) Must mention teacher/adults and children e.g. ".teacher is not in same age group as the children", "equation is for children not adults" "children and adults are different populations" "teacher will be taller" is B0 since no mention of children. "equation is only valid for children" is OK since "only" implies not suitable for adults Or Reference to different growth rates